Developing an Observational Rubric of Writing: Preliminary Reliability and Validity Evidence

Sinéad J. Harmey, Jerome V. D'Agostino, and Emily M. Rodgers

Sinéad Harmey, Department of Elementary and Early Childhood Education, Queens College, The City University of New York. Jerome D'Agostino, Department of Educational Studies, The Ohio State University; Emily Rodgers, Department of Teaching and Learning, The Ohio State University.

Sinéad Harmey is now at the International Literacy Centre, University College London Institute of Education, 20 Bedford Way, London, WC1H 0AL, U.K.

Correspondence concerning this article should be addressed to Sinéad Harmey, International Literacy Centre, University College London Institute of Education, 20 Bedford Way, London, WC1H 0AL, U.K.

Email: s.harmey@ucl.ac.uk

Abstract

The purpose of this paper was (1) to report on the design of the Early Writing Observational Rubric, designed to observe and describe change over time in the writing of children emerging into conventional literacy (ages 6-7) within an instructional setting and (2) to investigate the initial reliability and validity of the rubric. We used an extant data set that included 52 videos of writing instruction in Reading Recovery lessons (approximately 520 minutes) and pre- and post-intervention test data, for 24 students, taken at multiple time-points across a 20-week period. Dependent sample t-tests and HLM were used to ascertain if the rubric was sensitive to change over occasions. We also considered if the scores correlated with an external literacy measures. Findings suggest that the rubric has good initial reliability and validity and is a useful tool for researchers to observe and measure change over time in young children' writing as they write in an instructional setting; further validation work is required for use in other settings.

The contribution of early writing to early reading, and reciprocally, reading to writing, has long been theorized (Clay, 1982; Graham and Hebert, 2011; Shanahan, 2006). Yet, despite the well-accepted role of writing in early literacy growth, we have few tools to measure and describe change over time in early writing (Rowe and Wilson, 2015).

Assessment of early writing tends to involve proxy measures of writing ability (for example, name writing), or tests of encoding skills (for example, dictation). Curriculum based measures (cf. Ritchey, 2006) assess lower-level transcription skills like production and accuracy. Other measures emphasize conceptual changes in writing like concept of word (see Puranik and Lonigan, 2014). While potentially useful as screening tools, these measures may be less useful to inform writing instruction because they isolate and separate component writing skills and neglect to consider the orchestration of the act of composing and writing down a message. As Rowe and Wilson (2015) note, these tools are not very useful for classroom use because they require separate tests that differ from the kinds of writing that children produce in the classroom.

Instead, if one conceptualizes writing as a problem-solving activity in which the writer plans a message, generates it, and self-monitors the entire process (Flower and Hayes, 1977), then it makes sense to have a writing measure that can be used to systematically observe children as they engage in the act of writing within instructional settings. Such a tool could go beyond proxy measures and provide useful information about how students are composing a message and engaging in problem solving activities to write it down. This information would have the potential to not only provide information about writing progress but could also guide writing instruction for kindergarten and first grade students who are just emerging into conventional literacy. Rowe and Wilson (2015) produced such a tool, the Write Start! Writing Assessment, but its usefulness is limited to preschoolers between the ages of 2 and 5.

In this paper, we report on the development of the Early Writing Observational Rubric (EWOR) for slightly older children; those who have had at least one year of formal schooling but are still emerging as conventional writers. Drawing on the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014: 23) to frame the development of the EWOR, we used Standard 1, the overarching standard, to guide our work in that we sought to provide '*clear articulation of each intended test score interpretation for a specified use…and appropriate validity evidence.'* In the next section, we first review the literature that informs the conceptualization of the EWOR and then we describe the design of the rubric itself.

Theoretical frame

Yaden et al. (2000) identify several conceptual approaches used to study emergent writing. Outcome-based investigations view early writing as comprised of component skills, some of which are precursors to other skills. The work of Whitehurst and colleagues studying the emergent literacy of preschool children perhaps best exemplifies that approach (Storch and Whitehurst, 2002; Whitehurst and Lonigan, 1998; Whitehurst and Lonigan, 2001). Their model identifies oral language and code related precursors to conventional literacy. Storch and Whitehurst argue that while the influence of oral language abilities on reading diminishes (though does not disappear) once children start formal schooling, code-related skills consistently have a strong and direct impact on literacy development. In kindergarten, these code-related skills include these print principles: naming letters, identifying letters and sounds, differentiating print, knowing print components, segmenting words and segmenting sentences.

Wagner et al. (2011) take a similar approach in their model of the development of written language. Their analysis led to a five-factor model of writing development for first- through fourth-grade that includes: macro-organization of ideas (text level), productivity (word level;

number and diversity of words used) complexity, (sentence level; mean length of T-unit and syntactic density), spelling and punctuation, and handwriting fluency.

Both Whitehurst and colleagues' model of emergent literacy development and Wagner and colleagues' models of written language development provide important information about precursors to conventional literacy. An inherent shortcoming of models of literacy processes, however, is that often they do not provide information about how component skills might qualitatively change over time, nor do they provide a useful way to measure such change in instructional settings. As such, a sound observational rubric of early writing ought to not only reference component skills, but it should also describe how they change over time. In the sections that follow we describe our understanding of how emerging writing skills change over time and, in doing so provide rationales for the specific categories in our observational rubric. In the second part of the paper, we report on the evidence we collected in a study to examine the initial reliability and validity of the scores derived from the EWOR.

Development can be characterized as change over time

We conceptualize development as change over time in how children engage with a task (Thelen & Smith, 1994) and that learning occurs not only in independent settings when a child engages in a task alone, but also in interactions with more expert others (Vygotsky, 1987). From this perspective, there is value in observing a child's performance while engaged in an activity with a more expert other because these observations tell us something about a child's proximal performance; what the child can do with help (Vygotsky, 1987). Moreover, such information about what a child can do with help, better informs instruction in that teaching can be pitched at what the child is learning how to do, rather than what the child has already learned (Wood, 1998).

Our perspective on learning leads us to conceptualize effective assessment as the direct observation of children in the act of writing while working alongside a more expert other, who provides just enough help for the child to complete a task that would otherwise be too difficult to complete independently. As such, the rubric is designed to rate the child's control of the task and, at the same time, the degree of the teacher's participation

In terms of early writing, we are informed by literacy processing theory (Clay, 2001; for an explanation see Doyle, 2013), a theory that proposes two hallmarks of young children's growing control of reading and writing. One hallmark is that young children become more proficient over time in using the sources of information that they use to problem solve while reading or writing, and the other hallmark is that they become more proficient in the problem-solving actions they take. Within this context, our understanding of writing leads us to observe two features of the child's writing to assess their development: change in the sources of information used while writing, and change in what the child does to problem-solve while writing.

These two features of early writing development, change over time in sources of information used and change over time in problem-solving actions taken, comprise the two elements of the EWOR. Next, we identify the specific sources of information and problem-solving actions that comprise the EWOR and provide theoretical and empirical rationales for their inclusion.

Change over time in sources of information used while writing

What are the sources of information that young writers learn to use? Clay (2001) drawing from Rumelhart's (1994) notion of knowledge sources, identified several sources of information

that readers and writers use including oral language, letters, words, letter-sound relationships, phonological and orthographic coding, and rules about directionality of print.

***Using oral language to compose*.** Hayes (2011) proposed that there is a developmental trajectory that characterizes the act of composition in writing with children composing statements with one fixed topic and then moving towards elaborating about a topic. Hill (2011) described how young children must pay conscious attention to the structure of the message they write and that this involves choices about semantics and syntax in addition to print conventions and spelling (p.171). Oral language skills are related to writing ability (cf. Arfé, Dockrell and DeBernardi, 2016; Juel, 1986; Shanahan and Lomax, 1988). In a recent study, Kim and Schatschneider (2017) found that discourse level oral language skills had a substantial effect on writing quality over and above handwriting or spelling ability (p.43).

***Using orthographic information*.** Orthography is the spelling system of a language. Learning how to use orthographic information is particularly important in English, as it is an irregular language and only some words can be written accurately using only phonology (Ehri and Wilce, 1987; Ehri et al, 2009). Ehri (1989) defines this body of information as knowledge about the spelling system, memory for specific words and knowledge of letters to write them. Thus, being able to use orthographic information to spell words is an important, demanding, and complex evolving process. Indeed, Ehri and Wilce (1987) found that learning to spell contributed to superior results in word reading. Foorman and Petscher (2010) found that spelling ability had a positive relationship to text reading and comprehension. In a recent longitudinal study, Ouellette and Sénéchal (2017) found that invented spelling in kindergarten predicted later reading ability and mediated the connection between phonological awareness and early reading.

Using letter-sound relationships. Phonological awareness, an underlying process involved in reading and writing, refers to the skills involved in attending to, thinking about and manipulating the sound structures of words (Scarborough and Brady, 2002). Being able to use this knowledge and link this knowledge to a corresponding letter is important. It has been well established that phonemic awareness and letter-sound knowledge are reliable long-term predictors of learning to read and write (cf. Bourke, 2014; Ehri, Satlow and Gaskins, 2009; Hulme et al., 2012). Letter-sound knowledge is one of the best predictors of future decoding ability (cf. Hulme et al., 2012; Piasta, Phillips, Williams, Bowles and Anthony, 2016). Indeed, in a recent best-evidence synthesis Weiser and Mathes (2011) found that instruction in the use of letter-sound relationships and orthographic information improved spelling, phonemic awareness, decoding, and general writing ability, and indeed later reading ability.

Using a writing vocabulary. Production of written words is considered an indicator of general writing performance in many curriculum-based measures (Clay, 2013; Gansle, Noell, VanDerHeyden, Naquin and Slider, 2002; Ritchey, 2006). McCutchen (2006) stated that fluent text production is supported by mastery of lower level transcription processes which allows the writer to engage in higher level processes such as planning and reviewing (p.126).

Using print knowledge. Children's knowledge of the functions and conventions of print are important because they are related to the development of skills in both emergent and later literacy development (Puranik and Lonigan, 2014). As children’s literacy processing changes over time, Clay (2001) suggested that they exhibit a gradually more refined conception of how print works, from knowing merely that print represents a message to understanding that a collection of letters represents a word and each word is separated by spaces. A small yet robust body of research has demonstrated the important role of concept of word in overall literacy

development (Mesmer and Williams, 2015). The emergence of spacing and grouping together of words in print is regarded as an important hallmark in children's literacy development in that it demonstrates a developing awareness of one-to-one correspondence between text and spoken language (Ferreiro and Teberosky, 1982). Finally, print knowledge also includes control of the rule-governed directional properties of print (Ferreiro and Teberosky, 1982; Justice, Bowles and Skibbe, 2006).

Change over time in problem-solving actions while writing

Children must all become more proficient in the problem-solving actions they take while writing (Clay, 2013). These in-the-head activities of searching for more information and monitoring the accuracy of the text written are inferred by observations of particular behaviors like rereading to monitor the accuracy of what is already written (Chanquoy, 2009) or to generate new content or the next letter or word (van den Burgh, Rijlaardam and van Steendam, 2016), or self-correcting (cf. Fitzgerald, 1987). Self-correcting, revising or editing has been well considered in extant literature on models of writing development (cf. Bereiter and Scardamalia, 1987; Flowers and Hayes, 1977). It has been argued that students with difficulties in writing are less able to fluently control and monitor the writing process and engage in the necessary task of revision while writing (Graham and Harris, 1993: 170; Limpo, Alves and Fidalgo, 2013).

Transcription or handwriting fluency has also been found to be positively associated with later writing quality (Graham et al, 1997; Limpo, Alves and Connelly, 2017; Kim, Gatlin, Al Otaiba and Wanzek, 2017;). We suggest, therefore, that is also important to observe the speed with which children write their messages. Thus, as we shall outline in the section which follows, our rubric contains strands that focus on observable behaviours like rereading to monitor the accuracy of what is written, rereading to generate a new word, self-correcting, and transcription

fluency; all of which might indicate that young writers are engaging in these problem-solving activities.

Contents of the EWOR: Rating scale and items

The EWOR was designed for use while observing a child writing a message in an instructional setting (for example, one-to-one tutoring). Its design was an iterative process, theoretically informed by literacy processing theory as described in the previous section, with descriptive categories grouped per (1) what sources of information writers learn to use, and (2) what writers learn how to do as they write. In this section, we describe the EWOR rating scale (see Figure 1), then we describe each item of the rubric by providing examples of how each item is operationalized.

[Figure 1 about here]

Ratings: Using the 0-3 scale

The ratings of 0-3 represent the degree of control a child exercised when the action was observed. As such, and in line with our theoretical orientation to learning, each score also takes in to account the degree of teacher support that was needed to support the child to carry out the action (Vygotsky, 1987). For all items, a rating of zero indicates that the child was not observed to initiate an observable behavior (for example, never self-corrected), did not use a source of knowledge, or that the teacher provided the information or action needed to write the next part of the message. For example, the teacher wrote the word for the child. A rating of one indicates that the child was observed to either slowly initiate or needed high support to engage in a writing behavior and that this occurred on only one occasion. For example, a rating of one would be given if a child was observed to self-correct only one error. A hallmark of ratings of one is that

the production was slow and rarely occurred without some form of teacher support like prompting.

A rating of two indicates that the child was observed to initiate an observable behavior with some independence, on some occasions, and with minimal intervention from the teacher. For example, a rating of two would be given to a child who self-corrected on most occasions.

The highest rating of three would indicate that the child was observed to be in control of a certain facet of writing and almost always initiated the behavior or demonstrated efficiency and control without teacher intervention. For example, a rating of three would be given if a child self-corrected all errors.

With ratings on each item ranged from 0 to 3, the maximum sub-total score for the Using section of the rubric is 18 and the maximum sub-total score for the Doing section of the rubric is 12. The maximum total score possible, therefore, is 30. In the next section, we describe each item of the EWOR and provide examples of how the rating scale is used.

Section 1: Using sources of information

Using language to compose. The first item of the rubric describes the child's use of language to compose a message. After observing the child composing a message with teacher support, the rater considers the degree to which the child exhibited control of the composition process and whether they needed high or low support to do so. For example, a rating of zero would be assigned if the teacher told the child what to write whereas a rating of three would be assigned if the child composed without help.

Using orthographic information. The second item is used to rate the degree to which the child exhibits orthographic awareness. The rater considered if the child did not demonstrate any awareness of orthographic features of words, some awareness with prompting, awareness of

most features with minimal help, or awareness of all features with no help. We provide an example below to demonstrate the difference between a rating of zero and a rating of one;

Example;	The child is about to write the word ‘like’, writes ‘l-i’, and pauses.
Rating of zero:	The teacher takes the pen and says ‘let me write the next two letters for you they are ‘k’ and ‘e’. In this example, the child was not observed to exhibit any awareness of the orthographic features of the word.
Rating of one:	The teacher prompts the child ‘the word is like – think about how you would write that word to make it look right’. The child then writes the letters ‘k’ and ‘e’.

Using letter-sound knowledge. The next item concerns the extent to which the child uses knowledge about letter-sound relationships to both hear and record sounds in words. This item captures both hearing the smallest units of speech (phonemes) and then recording the corresponding grapheme. A rating of zero indicates the child needed support to both slowly articulate the word, hear sounds, and record letters. A rating of one is assigned when the child can hear and record dominant consonants with help. The highest rating indicates that the child heard and recorded sounds accurately without assistance.

Using writing vocabulary. This item describes the numbers of words that the child wrote independently and fluently while writing the short message. For this item in the rubric the observer must note if the child wrote any, some, or all words without help.

Using print knowledge. On the EWOR rubric this construct, print knowledge, is divided into two items; concept of word and directionality.

Concept of word. For this item, the observer must note whether the child placed spaces between words and whether the child needed constant, some, minimal, or no support to do so. Concept of word might be demonstrated by the child by, when faced with the end of the line, keeping the letters together in one unit (a word). We did not include this as the phenomenon might not be consistently be observed in writing short messages. In contrast, once a child moves beyond writing just a word the need to space words is ever present.

Directionality. When writing, letters, words and text are written in a certain order. For this item, the rater considers whether the child; needed constant support to move left to right (a rating of zero), demonstrates some control of directional movement (a rating of one), moved left to right but needed help at the end of line (a rating of two), or moved left to right and, when faced with no space at the end of the line, returned back to the left-hand side of the page to continue writing (a rating of three).

Section 2: Doing

We designed the rubric to include four categories to describe observable behaviors that imply that the child is engaged in these problem-solving activities: searching for more information, monitoring the accuracy of the message, revising if necessary, and working fluently (meaning, working with ease and speed). These processes are 'in-the-head' activities (Clay, 2001) but we theorized that some observable behaviors (like rereading) might infer problem-solving on the part of the child.

Rereading as if to search for more information. For this item, the rater considers whether the child initiated rereading as if to search for more information or help, to generate new content (van den Burgh et al., 2016) or write the next letter. An example of this would be if the child was writing the sentence '*I like cats'*. The child writes '*I like ca-*' then stops, returns to the

beginning of the sentence up to the point where he or she had stopped writing and then writes the next letter; in this case, *t* for cat. A rating of zero is assigned if this behavior was never observed, a rating of one is assigned if rarely observed, a rating of two is assigned if sometimes observed, and a rating of three is assigned if almost always observed. We cannot know what source of information the child was searching for because searching is a cognitive activity; we can only infer from the observable behavior that the child is conducting a search for more information to write the next letter.

Rereading for accuracy. For this item, the rater considers if the child initiated rereading as if to monitor the accuracy of what he or she wrote (cf. Chanquoy, 2009). An example of this would be if the child wrote the sentence '*I like cats'* and spontaneously reread the sentence as if to monitor whether it was written accurately. The difference between Rereading as if to Search for More Information and Rereading for Accuracy, is that the former takes place as the child is writing a word. By contrast, Rereading for Accuracy occurs after a word or the entire message is finished. Like the previous item, ratings of zero, one, two or three correspond with the behavior being never, rarely, sometimes, or almost always observed.

Self-correcting. Drawing on Fitzgerald (1987) we defined a self-correction as a writer's actions to identify a discrepancy between the intended and instantiated text, diagnosis of a solution, and correction of the error (p.484). Using the rubric, the rater considers whether the child self-corrected and, if so, how often this behavior was observed and the degree of fluency with which the child executed the self-correction.

Fluency. Given that fluency of text production is so important (McCutchen, 2006) the final item in this rubric, refers to the speed that the child wrote letters, words, and text. For this

item, the rater must consider whether the child's transcription was slow and labored, generally slow, mostly fast and fluent, or always fast and fluent.

In this section, we have described the design to establish the domain that would be measured by the EWOR. In the next section, we report on a study we conducted to establish the initial reliability and validity of the rubric.

Purpose of the study

Our primary purpose in designing the rubric was to have a tool that could be used in a research context to provide descriptive information about change over time in the writing of children who are nearing the end of the emerging phase and approaching conventional writing, as they write in an instructional context. In order ensure that the inferences drawn from the scores reflect the degree of writing proficiency of the student, we examined two conventional facets of reliability, inter-rater agreement and internal consistency, and two indices of validity, whether the scores changed over time and convergent-discriminant validity.

Study context and rationale

We applied the rubric to videos of writing instruction that occurred in the context of Reading Recovery (RR), a short-term early literacy intervention. The instructional setting was appropriate for three reasons. One, children were engaged in daily, one-to-one individualized instruction that included a writing component. Two, instruction was similar, though not scripted, across lessons. Teachers use similar instructional procedures in the writing component of a 30-minute lesson (for example, using Elkonin boxes to support the child to hear and record sounds in words). Finally, the writing component of the RR lesson follows a similar format for formulating a message in that; (1) the teacher engages the child in a genuine conversation and the topic is not controlled, (2) the conversation might revolve around something that interests the

child; a book they read, or something they did at home, and (3) when the conversation produces a likely utterance that the child could write with support the teacher helps the child to formulate the message to write (Clay, 2001: 27; Clay, 2005: 55). It should be noted that while the topic is not controlled (the children are not told what to write), the conversation does have the potential to have high teacher input if the child finds it difficult to orally formulate a short sentence. Clay (2001: 27) described the teacher's goal in these conversations is to increase initiation by the child.

Method

We used an extant data set, originally collected for another study that contained test data and videos of teaching.

Student-teacher dyads

The videos of teaching were collected from 24 student-teacher dyads. In Autumn of First grade, teachers identified a cohort of children as among the lowest-achieving in terms of literacy learning. This cohort was screened using the Observation Survey of Early Literacy Achievement [OSELA] (Clay, 2013) and the lowest achieving students were identified and placed in RR to participate in the literacy intervention. The students in the data set were, indeed, experiencing significant difficulties in literacy learning but were not eligible for special education. The average pre-intervention OSELA total score (D'Agostino, 2012)(M = 361.8, SD = 40.6) placed this cohort in the 9th percentile compared to a national random sample of their peers, placing them at risk of later difficulties in literacy learning.

The group comprised 14 girls and 10 boys. Most students were Black (n =16), the rest were White (n = 8), one quarter (n = 6) spoke English as an additional language and all were entitled to free school lunches. The average age of students who received a RR intervention in

Autumn of 2014 – 2015 was 6 years and 5 months (D'Agostino and Harmey, 2015). The students went to 22 schools from one urban school district in a mid-west city in the United States of America. Most children in these schools were considered economically disadvantaged and racially diverse.

There were 22 teachers (2 teachers taught 2 students each), all of whom were female. In general, they were experienced elementary school teachers with an average of 17.45 (*SD* = 6.2) years of experience teaching. 19 teachers were in the training year of the RR professional development certification and 3 teachers had trained the previous year. Five teachers were Black and 17 were White.

Measure

Besides the EWOR scores, we collected OSELA (Clay, 2013) scores on the students to conduct the convergent-discriminant validity analysis. The OSELA was particularly useful in conducting the analysis because it contains two tasks that are more closely aligned with the EWOR, and four tasks that we assumed were less conceptually associated with the EWOR. The two tasks that we predicted would yield scores that would converge with the EWOR scores were Writing Vocabulary (WV) and Hearing and Recording Sounds in Words (HRSW). The former is a timed task where the student writes as many words as possible in ten minutes and receives credit for each word spelled correctly. The latter is a dictation task where the student encodes a dictated sentence and receives credit for each phoneme correctly recorded. We expected the other four OSELA measures, Letter Identification (LI),), Concepts About Print (CAP), the Ohio Word Test, and Text Reading, to be less associated with EWOR scores because they are more associated with reading rather than writing skills.

According to the National Center on Response to Intervention [NCRTI] (2016), the OSELA is one of the highest rated literacy screening tools for literacy and has convincing evidence in terms of reliability, validity, generalizability, and disaggregated data for diverse populations. The NCRTI also found that the OSELA had broad generalizability. Split-half reliability coefficient value for the OSELA was .89. The alpha coefficient was .87. The OSELA had content, construct, and predictive validity with values above .70 and the OSELA was found to have correlations with other measures of early literacy (e.g., the Slosson Oral Reading Test; Slosson and Nicholson, 2002) (D'Agostino, 2012).

Sources of data

For the 24 student-teacher dyads there were 52 videos of the writing segment of the child's Reading Recovery lesson (20 students videoed at two times points and 4 students videoed at three time points during the series of lessons). For each child, all writing event videos were rated using the EWOR. The total instructional time of the videos was approximately 520 minutes.

Procedure

Initial construction of the rubric. We reviewed extant tests to consider constructs and items in early writing measures. We constructed the items, framed theoretically by literacy processing theory (Clay's notion of using sources of information and doing). The construction of the EWOR was an iterative process that involved writing items, piloting the rubric, refining language, and using the rubric again.

Piloting. We conducted a pilot study by rating 15 videos of a writing segment of a RR lesson. The mean total score was 14.53 (*SD* = 5.43) and was slightly negatively skewed in a histogram of score frequencies. Internal consistency of the scores assigned was calculated and

the alpha coefficient was .87. During the process of the piloting phase, language in the rubric was further clarified and titles of rubric items were refined. Two other raters were then trained by the first author to use the rubric.

Data Analysis Plan

Following the construction and piloting of the rubric we set out to collect evidence about the reliability and validity of the rubric.

Reliability. To examine the degree of inter-rater reliability, we compute Cohen's kappa and the intra-class correlation (ICC). Cohen's kappa was calculated to measure the observed level of agreement between the first author and two other raters, both graduate students in reading and literacy, correcting for agreement that would be expected by chance (Hallgren, 2012). As kappa statistics are more suitable for nominal ratings, an ICC was calculated as it is more suited if IRR is obtained with ordinal or ratio variables, and incorporates the magnitude of disagreement into the estimate (Hallgren, 2012).

Besides examining the reliability of raters, we evaluated the internal consistency of rubric scores. To do this, we calculated descriptive statistics and measures of central tendency to see if the rubric was sensitive to intra-individual variability. Internal consistency, as represented by the alpha-coefficient variable, describes the estimates of reliability based on the average correlation among items in a test (Nunally and Bernstein, 1994).

We ascertained the internal consistency, or the general agreement between the items of the EWOR. An inter-item correlation matrix was created and alpha coefficient was calculated (Tavakol and Dennick, 2011). In addition to the alpha coefficient, Pearson's correlation coefficients between (1) each item, (2) between items and total scores and sub-total scores were calculated.

Validity. The AERA, APA, and NCME (2014) outlined that the types of validity evidence that need to be collected is entirely dependent on the context in which the test is used. In this study, the EWOR was designed for research purposes to describe change over time in early writing. We ascertained if the scores indeed reflected this construct by considering if the scores were sensitive to changes and were more correlated with extant measures that were conceptually similar to writing and less correlated with extant measures that were less conceptually related to the construct.

To ascertain if the EWOR was sensitive to change, we used dependent sample t-tests of the scores obtained at first and last observation to examine if there were significant differences between each occasion. To further explore whether the scores reflected writing development we used hierarchical linear modeling (Raudenbush and Bryk, 2002) to estimate rates of growth over time. Finally, we considered if the scores obtained correlated with other extant measures of reading and writing by using the students' pre- and post-intervention OSELA (Clay, 2013) tasks and total scores with item and total scores of the first and final EWOR ratings. We correlated the EWOR scores with pretest and posttest OSELA task and Total scores to examine convergent and discriminant validity.

Results

Reliability

Inter-rater reliability. The first author rated all videos and these ratings were compared to two other raters who rated 10% of the videos each (n = 11). Two statistical variables were calculated to assess IRR for ratings on each item in the rubric. The kappa values ranged between $\kappa = .62$ to $\kappa = 1.0$ which indicated that agreement ranged from substantial to perfect (see Table 1). The ICC values ranged between .78 and 1.0 ($p < .05$).

[Table 1 about here]

Internal consistency. We calculated measures of central tendency and variability, internal consistency, and inter-item correlations to evaluate if the scores derived from the EWOR were reliable or reflected writing achievement at a given time. Having rated each writing observation, we first calculated measures of central tendency and variability (see Table 2). There was more variability in the ratings of the using section of the rubric, as represented by the standard deviations, at the final observation compared to the first observation. The one exception was the standard deviations for the items that measured concept of word and directionality, where there was little variation in ratings at the final observation.

[Table 2 about here]

The alpha-coefficient for the EWOR scores for the first observation was .76, which is acceptable (Nunally and Bernstein, 1994). Eliminating the composition item would have raised the alpha co-efficient by .03. Eliminating the Rereading as if to Search for More Information item would have reduced the alpha coefficient by .07. None of the other items would have raised or reduced the alpha coefficient by any degree greater than those described. The alpha coefficient for the EWOR scores for the final observation was .82. Deleting any of the items would not have raised the alpha co-efficient; rather, it would have either have kept the value the same or reduced it by up to .05.

Correlations were calculated between items, sub-total scores, and total scores. This was calculated at the first (see Table 3) and final observation. We also calculated correlations between OSELA (Clay, 2013) writing tasks and total scores, which will be reported in the next section. For the first observation, use of language to compose a message did not have a

statistically significant correlation with either the using sub-total, $r_{(48)} = .31$, or the total score, $r_{(48)} = .21$. This pattern was noted in the piloting of the rubric.

[Table 3 about here]

By the final observation (see Table 4) composing did have a statistically significant correlation with both the using sub-total score ($r_{(48)} = .74$, $p < .01$), and the total score, ($r_{(48)} = .68$, $p < .01$).

[Table 4 about here]

The correlations between the items in the using section of the rubric and the using sub-total score were greater than the correlations with the doing sub-total score, at both the first and final observations. This pattern was reciprocated for the items in the doing section of the rubric, in that they were more highly correlated with the doing sub-total score at both the first and final observations. At the first observation, the correlations between all items and the total score were positive and statistically significant (except for self-correction, $r_{(48)} = .3$). At the final observation, the correlations between all items and the total score were positive and statistically significant (except for directionality, $r_{(48)} = .28$).

Validity

Change over time in EWOR scores. If the scores derived from the EWOR reflected change over time in writing, the scores obtained should change as children became more proficient. In other words, if children were observed to become more independent in composing, spelling, spacing, and monitoring the accuracy of their message their scores should increase. We tested this in two ways; we conducted a dependent samples T-test and used HLM. The dependent samples t-test was used to compare each student's first and final ratings for the using sub-total, doing sub-total, and total score. Results indicated that total scores were significantly higher at the

final observation (M = 18.04, SD = 3.78) compared to the first observation (M = 15.41, SD = 3.09), as indicated by a significant t-test, $t_{(23)} = 3.72$, 95% CI [1.16, 4.09], $p < .001$. Using sub-total scores were significantly higher at the final observation (M = 12.71, SD = 2.22) compared to the first observation (M = 10.79, SD = 1.72), as indicated by a significant t-test, $t_{(23)} = 4.18$, 95% CI[.97, 2.86], $p < .001$. Doing sub-total scores were significantly higher at the final observation (M = 5.33, SD = 1.86) compared to the first observation (M = 4.62, SD = 1.74), as indicated by a significant t-test, $t_{(23)}$, 2.06, 95% CI[-.01, 1.42] $p < .05$.

To further explore whether the scores reflected change over time in writing we conducted HLM (Raudenbush and Bryk, 2002) to estimate a rate of growth over time for each item and for the Doing sub-total, the Using sub-total, and the EWOR total score. Scores were nested within time (t) and children (i), with the time variable (coded '0' for the first observation, and '1' for the final observation) entered as a level-1 predictor. No child-specific predictors were entered at Level-2 (hence, an unconditional model). The equations at Levels 1 and 2 were:

$$\text{Level 1:} \quad Y_{ti} = \pi_{01} + \pi_{1i}(\text{time}) + e_{ti}$$

$$\text{Level 2:} \quad \pi_{01} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10} + r_{1i},$$

where β_{10} indicated the average slope and β_{00} reflected the average first score, while $+ r_{0i}$ and r_{1i} were the intercept and slope residuals, respectively, for each student.

Table 5 provides the average intercept and slopes for the items, sub-totals, and total score, along with the standard errors (SE) for each coefficient, and the results of t-tests that examined the hypotheses that each coefficient equaled zero. The coefficients in the table allow one to predict the average initial status and average rate of change from first to final observation for each item, sub-total, or total score. For example, the average first observation score for the

Using sub-total was 10.58, and students grew, on average, 1.74 score points from first to final observation for a predicted final observation score of 12.32. Both the average initial status and slopes were significantly greater than zero.

[Table 5 about here]

It can be seen from the table that three of the six Using items (Composing, Use of Orthographic Information, Use of Letter-Sound Relationships) and one of the five Doing items (Fluency) were sensitive to change over time as evidenced by the significant growth rates for those items. The average slopes for the Using and Doing sub-totals also were significant even though half of the Using and all but one of the Doing items were not statistically greater than zero. Note, however, that some growth occurred on all items, and although not statistically significant in all cases, cumulated across items to render significant sub-total growth rates. Because the EWOR total score was the combination of the two sub-totals, it also demonstrated sensitivity to change from first to final observation across students.

If the EWOR scores reflected writing proficiency, then the scores derived would correlate with scores from writing tasks and would not correlate with scores from non-writing tasks. Given that all students were tested at the beginning and end of the RR intervention using the OSELA (Clay, 2013) we conducted bivariate correlations between pre- and post-intervention OSELA scores and first and last observation EWOR ratings. We hypothesized that there would be stronger correlations between scores on the writing tasks of the OSELA (WV and HRSW) and the OSELA total score. In contrast, we hypothesized that the reading related tasks (for example, Text Reading Level) would not be as strongly correlated.

We found that, indeed, that there was a predictable discriminant pattern with no statistically significant correlations between reading tasks (Text Reading Level and Concepts

About Print) and the items and total score of the EWOR. For example, the correlation between EWOR first observation total score and pre- intervention Text Reading Level was not statistically significant ($r = .14$, $p = .51$). There were some scattered 'hit or miss' positive correlations between scores on some tasks of the OSELA and items on the EWOR. For example, there were statistically significant correlations between scores on the pre-intervention Letter Identification and the first observation EWOR Doing Sub-total scores ($r = .40$, $p < .05$), post-intervention Ohio Word Test scores and last observation EWOR scores for use of orthographic information ($r = .43$, $p < .05$).

In contrast, post-intervention WV scores were correlated with EWOR first observation use of letter-sound information scores ($r = .39$, $p < .05$) and first observation EWOR fluency scores ($r = .45$, $p < .05$), and EWOR doing sub-total scores ($r = .41$, $p < .05$) (see Table 3). There was an even more consistent pattern of convergence between post-intervention WV scores and EWOR final observation scores (see table 4). Post-intervention WV scores were positively correlated with final observation scores on the EWOR use of orthographic information item ($r = .47$, $p < .05$), EWOR self-correction item ($r = .58$, $p < .01$), EWOR fluency item ($r = .42$, $p < .05$), EWOR doing sub-total ($r = .48$, $p < .05$), and EWOR total score ($r = .47$, $p < .05$). First observation EWOR fluency was positively correlated with entry HRSW ($r = .39$, $p < .05$). Final observation EWOR reading for accuracy was positively correlated with exit HRSW ($r = .51$, $p < .05$).

Discussion

The purpose of this study was to (1) report on the design of the EWOR and (2) investigate the initial reliability and validity of the scores obtained on the EWOR. In this section, we provide a discussion about the design of and our findings about the initial reliability and

validity of the EWOR, we acknowledge the limitations of this study, and conclude by proposing avenues for future research.

Design of the Observational Writing Rubric

The EWOR was designed as a tool to observe children's use of sources of information and revision behaviors as they wrote in an instructional context. Like Camp (2012), we frame writing as multi-dimensional and suggest that development is dependent on instruction (Glasswell, 1999). Theoretically, we framed writing as a process in which a child uses multiple sources of information to problem-solve the task of producing a meaningful message (cf. Harmey, 2015). This implies that to gather data about changes in early writing, a researcher needs an observational tool to 'capture the learner at work' (Clay, 2001; 82). While we acknowledge that control of these dimensions of early writing (such as use of oral language to compose a message, use of letter-sound relationships, use of orthographic information, adherence to the directional rules of print, and revision) can be inferred by analyzing what a child produces (the written product) we argue that they can be directly observed as the child writes in an instructional context thus greatly reducing the reliance on inferring what happened. Such an alternative stance demands observation of the origins of self-regulation of the writing processes (Diaz, Neal, and Amaya-William, 1990) within the context that it occurs, as evidenced by observable writing behaviors. To our knowledge, no such measure exists. Thus, the EWOR provides a viable alternative measure that can be used to systematically observe young writers at work within an instructional setting, and can capture change over time in control of observable writing behaviors.

It should be noted that the contents of the EWOR rests on a particular conceptualization of early writing development as we outlined earlier in this paper's theoretical frame. Thus, we

anticipate that those who share the same views of learning and of early writing will find the EWOR to be more useful and informative than those who hold a different view of writing development. Some for example may only be interested in the product, not the composition and production of the message and some may only be interested in measuring what the student can do independently, not with teacher help.

Moreover, the EWOR is necessarily dense in its contents, dense because early writing development is so complex. As such, we imagine that it may take some practice for novices to the EWOR to learn what to observe and how to score it.

The initial reliability and validity of the EWOR

Reliability. We established the initial reliability and validity of the scores obtained from the EWOR. In terms of inter-rater reliability, we achieved substantial agreement between the first author and two other raters. We have found few early writing rubrics that provide evidence of reliability and validity of the scores obtained beyond inter-rater reliability (e.g. Calkins, 2013; Watanbe and Hall-Kenyon, 2011). This is even though Nunally (1978) recommended that further investigations of reliability should always be conducted in the development of any new measures.

Our results using measures of central tendency and variability indicated that the scores obtained reflected intra-individual variability. There was, however, a ceiling effect for the items that measured print knowledge (concept of word and directionality). While print knowledge certainly has an important role to play in emergent literacy (Justice et al., 2006), it may have been that the children in our study, who were moving towards conventional writing, already controlled this source of knowledge prior to the intervention.

We examined the internal consistency of the items on the EWOR and our results demonstrated that the items in the EWOR were, importantly, measuring constructs that constituted part of the domain of early writing (Tavakol and Dennick, 2011). The positive correlations between the items in the EWOR indicated that the items are not so high that one might suspect that they are measuring the same thing. The alpha coefficient values were acceptable (Tavakol and Dennick, 2011) and, thus provided a good estimate of the reliability of the scores derived from the EWOR (Nunally, 1978).

Validity. Validity refers to the inferences one can make from test scores (AERA, APA and NCME, 2014). Our results demonstrated that there were statistically significant differences between scores over time and we could detect positive rates of change using HLM, some of which were statistically significant. Taken together, these results provide an initial source of evidence towards the establishment of initial validity of the scores as reflective of writing development.

Another source of evidence about the validity of the scores derived are the findings that the scores derived from items on the EWOR correlated with both pre- and post-intervention WV and some HRSW (Clay, 2013) scores. These results are promising given that the WV task captures writing skills including but not limited to transcription fluency, word production, spelling, writing vocabulary, and use of orthographic information and is similar to extensively used curriculum-based measures (number of words written in a given time) (cf. Ritchey, 2006). Taken together, these results provide further evidence of the validity of the scores derived from the EWOR as a valid measure of writing.

Limitations

This study was not void of limitations. One limitation of this study was that we did not directly address or code the impact of the teacher's support. We acknowledge that the support provided by the teacher will directly affect the child's responses. We could have coded teacher behavior. We do argue, however, that given the way the scale is operationalized that teacher support is implicitly acknowledged. Take, for example, the first item on the rubric that considers the child's control of the composition process. If a child scored zero it is noted that the teacher provided high support. In contrast, if a child scored two the teacher provided less support. Thus, the items provide a representation of the degree of control of an action a child exhibited or regulated.

Another limitation of this study is that the reliability and validity evidence was gathered on a small and homogenous group of children limited to one instructional context and this limitation may present a challenge to replication. Although we maintain that the context was ideal for collecting initial reliability and validity evidence because the instructional approach was uniform and the students were at similar points in terms of literacy learning, the fact that the children were struggling learners meant that we were unable to detect much growth on some items. Considering the overall lack of growth for a number of sampled students, however, the EWOR items were still quite sensitive to change over time. As we continue to collect more reliability and validity analyses of the scores, it will be imperative that we do so with a more heterogeneous sample that considers the full range of student achievement and growth rates in more diverse instructional settings.

As stipulated in the *Standards for Educational and Psychological Testing* (AERA, APA and NCME, 2014), the onus rests with test users to collect the reliability and validity evidence to support the interpretations based on the specified purposes of a measure. Thus at present, the

utilization of the rubric is limited to the particular type of setting in this study: observing a student who is just emerging in to conventional literacy, composing and writing a message with teacher help. As use of the EWOR is replicated in different settings for different purposes, there will be a need to collect additional reliability and validity information. For example, the EWOR potentially could be used as a screening tool, a self-reflection tool by teachers, a professional development tool by coaches, or indeed as a formative assessment measure to guide learning. Different facets of validation evidence will be required to support the tool scores for each of those purposes, such as predictive validity evidence if used as a screener, or consequential validation evidence if used by teachers for self-reflection or in a formative assessment role. Our study was limited to collecting evidence to support the EWOR's development and use in a research setting.

Future directions

As a tool in development the process of establishing reliability and validity is ongoing and iterative. The AERA, APA, and NCME (2014: 12) described how new evidence could influence the interpretations that can be drawn from the scores and the conceptual framework of the test. A next step in our research agenda is to further validate the tool in different contexts with a more diverse population. As we develop the tool, one of our lines of research is to train literacy teachers to use the EWOR as a tool for self-reflection to support struggling writers. As new evidence becomes available, we will continue to evaluate the evidence about the reliability and validity of the EWOR and, indeed, its design. As it stands, it serves as a useful tool that researchers can use to observe and describe the writing development of young children as they emerge into conventional literacy.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (2014) *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Arfé B, Dockrell JE and DeBernardi B (2016) The effect of language specific factors on early written composition: the role of spelling, oral language and text generation skills in a shallow orthography. *Reading and Writing* 29(3): 501-527.

Bereiter C and Scardamalia M (1987) *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum.

Bourke L, Davies SJ, Sumner E, et al. (2014) Individual differences in the development of early writing skills: Testing the unique contribution of visuo-spatial working memory. *Reading and Writing* 27(2): 315-335.

Calkins L (2013) *Writing Pathways: Performance Assessments and Learning Progressions, Grades K-8.* Portsmouth, NH: Heinemann.

Camp H (2012) The psychology of writing assessment. *Assessing Writing,* 17(2): 92 - 105

Clay MM (1982) *What Did I Write?* Auckland, NZ: Heinemann.

Clay MM (2001) *Change Over Time in Children's Literacy Development*. Auckland, NZ: Heinemann Educational Books.

Clay MM (2005) *Literacy Lessons Designed for Individuals: Part 2 Teaching Procedures*. Portsmouth, NH: Heinemann.

Clay MM (2013) *An Observation Survey of Early Literacy Achievement*. Portsmouth, NH: Heinemann.

Chanquoy L (2009) Revision processes. In: Beard R, Myhill D, Riley J and Nystrand M (eds) *The Sage Handbook of Writing Development*. London: Sage, pp.80-97.

D'Agostino, JV (2012) Technical review committee confirms highest NCRTI ratings for Observation Survey of Early Literacy Achievement. *Journal of Reading Recovery* 11(2): 53-56.

D'Agostino JV and Harmey S (2015) Reading Recovery and Descubriendo la Lectura national report 2013-2014 (IDEC Rep. No. 2015-01). Columbus: The Ohio State University, International Data Evaluation Center.Diaz R, Neal C and Amaya-William M (1990). *The social origins of self-regulation*. In L.C. Moll (Ed.), Vygotsky and education: Instructional implications and applications of sociohistorical psychology. Cambridge: Cambridge University Press, p. 127 – 154.

Doyle MA (2013) Marie M. Clay's theoretical perspective: A literacy processing theory. In: Alvermann DE, Unrau N and Ruddell RB (eds) *Theoretical Models and Processes of Reading, Sixth Edition*. Newark, DE: International Reading Association, pp.636-656.

Ehri LC (1989) The development of spelling knowledge and its role in reading acquisition and reading disability. *Journal of Learning Disabilities* 22(6): 356-365.

Ehri LC and Wilce LS (1987) Does learning to spell help beginners learn to read words? *Reading Research Quarterly* 22(1): 47-65.

Ehri LC, Satlow E and Gaskins I (2009) Grapho-phonemic enrichment strengthens keyword analogy instruction for struggling young readers. *Reading & Writing Quarterly* 25: 162-191.

Ferreiro E and Teberosky A (1982) *Literacy Before Schooling*. Exeter, NH: Heinemann Educational Books.

Fitzgerald, J. (1987). Research on revision on writing. *Review of Educational Research* 4(57): 481-506.

Flower LS and Hayes JR (1977) Problem-Solving strategies and the writing process. *College English* 39(4): 449-461.

Foorman B and Petscher Y (2010). Development of spelling and differential relations to text reading in grades 3-12. *Assessment for Effective Intervention* 36(1): 7-20.

Gansle KA, Noell GH, VanDerHeyden AM, et al. (2002) Moving beyond total words written: The reliability, criterion validity, and time cost of alternative measures for curriculum-based measurement in writing. *School Psychology Review* 31(4): 477-497.

Glasswell K (1999). *The patterning of difference: Teachers and children constructing development in writing* (Unpublished doctoral dissertation). University of Auckland, New Zealand.

Graham S, Berninger VW and Abbott RD et al (1997) Role of mechanics in composing of elementary school students: A new methodological approach. *Journal of Educational Psychology* 89 (1): 170-182

Graham S and Harris KR (1993) Self-regulated strategy development: Helping students with learning problems develop as writers. *The Elementary School Journal* 94(2): 169-180.

Graham S and Hebert M (2011) Writing to read: A meta-analysis of the impact of writing and writing instruction on reading. *Harvard Educational Review* 81(4): 710-744. DOI: https://doi.org/10.17763/haer.81.4.t2k0m13756113566

Hallgren KA (2012) Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quantitative Methods Psychology* 8(1): 23-34. DOI: https://doi.org/10.20982/tqmp.08.1.p023

Harmey S (2015) *Change over time in children's co-constructed writing* (Doctoral dissertation). Department of Teaching and Learning, The Ohio State University, Columbus.

Hayes JR (2011) Kinds of knowledge-telling: Modelling early writing development. *Journal of Writing Research* 3(2): 73-92. DOI: 10.17239/jowr-2011.03.02.1.

Hill, S. (2011). Towards ecologically valid assessment in early literacy. *Early Child Development and Care,* 181(2), 165-180.

Hulme C, Bowyer-Crane C, Carroll JM, et al. (2012) The causal role of phoneme awareness and letter-sound knowledge in learning to read: Combining intervention studies with mediation analyses. *Psychological Science* 23(6): 572-577.

Juel, C (1988) Learning to read and write: a longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology 80*, 437-447

Justice LM, Bowles RP, Skibbe LE (2006) Measuring preschool attainment of print-concept knowledge: A study of typical and at-risk 3-to 5-year-old children using item response theory. *Language, Speech and Hearing Services in Schools* 37(3): 224-235.

Kim YSG, Gatlin B and Al Otaiba S and Wanzek J (2017) Theorization and an empirical investigation of the component-based and developmental text writing fluency construct. *Journal of Learning Disabilities.* Prepublished June, 9, 2017.

Kim YSG and Schatschneider C (2017) Expanding the developmental models of writing: A direct and indirect effects model. Journal of Educational Psychology 109(1), 35-50.

Limpo T, Alves RA and Connelly V (2017). Examining the transcription-writing link: Effects of handwriting fluency and spelling accuracy on writing performance via planning and translating in middle grades. *Learning and Individual Differences*, 53, 26-36.

Limpo T, Alves RA and Fidalgo R (2013) Children's high-level writing skills: Development of planning and revising and their contribution to writing quality. *British Journal of Educational Psychology* 84(2); 177-193.

McCutchen D (2006) Cognitive factors in the development of writing. In: McArthur CA, Graham S and Fitzgerald J (eds) *Handbook of Writing Research*. New York, NY: The Guildford Press, pp.115-130.

Mesmer HA and Williams TO (2015) Examining the role of syllable awareness in a model of concept of word: Findings from preschoolers. *Reading Research Quarterly* 50(4): 483-497.

National Center on Response to Intervention (2016) *Screening tools chart: Observation survey of early literacy achievement*. Available at: www.rti4success.org/observation-survey-early-literacy-achievement-reading.

Nunally JC (1978). *Psychometric theory*. New York: McGraw-Hill.

Nunally JC and Bernstein IH (1994) *Psychometric Theory*. New York, NY: McGraw-Hill.

Ouellette G and Sénéchal M (2017) Invented spelling in kindergarten as a predictor of reading and spelling in Grade 1: A new pathway to literacy, or just the same road, less known? *Developmental Psychology* 53(1): 77-88.

Piasta SB, Phillips BM, Williams JM, et al. (2016) Measuring young children's alphabet knowledge. *The Elementary School Journa*, 116(4); 524-548.

Puranik CS and Lonigan CJ (2014) Emergent writing in preschoolers: Preliminary evidence for a theoretical framwork. *Reading Research Quarterly* 49(4): 453-467.

Raudenbush SW and Bryk AS (2002) *Hierarchical Linear Models: Applications and Data Analysis Methods.* Thousand Oaks: Sage Publications.

Rumelhart DE (1994) Toward an interactive model of reading. In: Alvermann DE, Unrau N and Ruddell RB (eds) *Theoretical Models and Processes of Reading, Fourth Edition*. Newark, DE: International Reading Association, pp. 864 894.

Rowe DW and Wilson SJ (2015) The development of a descriptive measure of early childhood writing results from the Write Start! Writing Assessment. *Journal of Literacy Research* 47(2): 245-292.

Ritchey KD (2006) Learning to write: Progress-monitoring tools for beginning and at-risk writers. *Teaching Exceptional Children* 39(2): 22-26.

Scarborough, H. & Brady, S. (2002) Toward a common terminology for talking about speech and reading: A glossary of the 'phon' words and some related terms. *Journal of Literacy Research, 34* (3), 299-336.

Shanahan T (2006) Relations among oral language, reading, and writing development. In: Shanahan T, MacArthur CA, Graham and Fitzgerald J (eds) *The Handbook of Writing Research.* New York, NY: Guildford Press, pp.171-183.

Shanahan T and Lomax RG (1988) A developmental comparison of three theoretical models of the reading-writing relationship. Research in the Teaching of English 22(2), 196 – 212.

Slosson, R.L., & Nicholson, C.L. (2002). *Slosson oral reading test, third edition (SORT-R3).* East Aurora, NY: Slosson Educational Publications, Inc.

Storch SA and Whitehurst GJ (2002) Oral language and code-related precursors to reading: evidence from a longitudinal structural model. *Developmental Psychology* 38(6): 934-947. DOI: https://doi.org/10.1037/0012-1649.38.6.934

Tavakol M and Dennick R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education* 2: 53-55.

Thelen E and Smith LB (1994) *A dynamic systems approach to the development of cognition and action.* MIT Press

van den Burgh, Rijlaardam and van Steendam. (2016). Writing process theory. In MacArthur, C.A., Graham, S. and Fitzgerald, J. (Eds.), *The Handbook of Writing Research (pp. 57 – 69).* New York, NY: The Guildford Press.

Vygotsky LS (1987) Thinking and speech. In: Rieber R and Carton A (eds) *The Collected Works of L.S. Vygotsky, Volume 1*. New York, NY: Plenum Press, pp.39-285.

Wagner RK, Puranik CS, Foorman B, et al. (2011) Modeling the development of written language. *Reading and Writing* 24, 203-220.

Watanabe LM and Hall-Kenyon KM (2011) Improving young children's writing: The influence of story structure on kindergartners' writing complexity. *Literacy Research and Instruction* 50(4): 272-293.

Weiser BL and Mathes PG (2011) Using encoding instruction to improve the reading and spelling performances of elementary students at-risk for literacy difficulties: A best-evidence synthesis. *Review of Educational Research* 81(2): 170-200.

Whitehurst GJ and Lonigan CJ (1998) Child development and emergent literacy. *Child Development* 69(3): 848-872. DOI: https://doi.org/10.1111/j.1467-8624.1998.tb06247.x

Whitehurst GJ and Lonigan CJ (2001) Emergent literacy: Development from prereaders to readers. In Neuman SB and Dickinson DK (eds.) *Handbook of early literacy research.* New York: Guilford Press, pp. 11 – 29.

Wood D (1998) *How Children Think and Learn*. Oxford: Blackwell Press.

Yaden DB, Rowe DW and McGillivray L (2000) Emergent literacy: A matter (polyphony) of perspectives. In Kamil ML, Mosenthal PB, Pearson PD and Barr R (eds) *Handbook of reading research.* Mahwah, NJ: Erlbaum, pp. 425 – 454.

Figure 1. Observational Rubric of Writing

	Observation of Writing					
	Item	**Score of 0**	**Score of 1**	**Score of 2**	**Score of 3**	Total
USING	Use of Language to compose	Did not initiate/struggled to compose message without very high support or was told what to write.	Slow to initiate composition of a simple message. Needed high support to construct message.	Exhibited control of parts of the conversation and composition. With support expanded message.	In control of the conversation/had a message ready to write. Composing was fluent and was flexible to make changes on the run.	
	Use of Orthographic information	Did not demonstrate any awareness of orthographic features of words. Teacher contributed information.	Demonstrated some awareness of the orthographic features of words with prompting.	For many words, demonstrated some awareness of the orthographic features of words with minimal help.	Demonstrated awareness of the orthographic features of words and words were mostly spelled accurately and with efficiency	
	Use of Letter-sound knowledge	Did not initiate slow articulation of words. Needed support to say word slowly, hear, and record sounds.	With prompting, could say word slowly and hear and record some initial sounds and dominant consonants with support.	Initiated slow articulation and heard and recorded phonemes in words from beginning to end with minimal support.	Initiated slow analysis of words independently and accurately (sometimes using vocalization to break a word apart or silently).	
	Use of Writing Vocabulary	Did not write any words independently	Wrote one word independently. Process was slow. On all other occasions required support.	Wrote some words independently and with some speed with minimal support.	Wrote all words quickly, efficiently and independently without support.	
	Use of Print Knowledge	Did not initiate placing spaces between words and needed constant direction.	Sometimes initiated making spaces between words but still needed support.	Spaced words correctly with minimal intervention.	Put spaces between words efficiently and needed no reminders to attend to this.	
		Did not initiate movement from left to right and needed constant support.	Sometimes showed control of directional movement but still needed support.	Moved left to right with minimal intervention but needed reminder to go to a new line when out of space.	Moved left to right quickly and efficiently. Moved to a new line when needed and needed no reminders to attend to this.	
DOING	Rereading as if to seek help	Did not initiate rereading to seek help writing the next letter	Rarely initiated rereading to seek help to write the next letter.	Sometimes initiated rereading to seek help to write the next letter.	Almost always initiated rereading to seek help to write the next letter.	
	Rereading for accuracy	Did not initiate any rereading to check the accuracy of what was written.	Rarely initiated rereading to check the accuracy of what was written.	Sometimes reread to check that the message was accurate with minimal support.	Almost always reread to check accuracy in a fast efficient manner with no support.	
	Self-correcting	If error was made did not notice or correct it	If error was made, noticed and self-corrected on one occasion.	If errors were made, noticed and self-corrected with some speed on most occasions.	When errors were made was fast to self-correct or wrote independently without error.	
	Fluency	Writing was slow and laboured. Required high support to form letters or words.	Writing was generally slow but for known words or letters but pace picked up.	Writing was mostly fast and fluent but faltered over formation of some letters or words.	Transcription was fast and fluent.	
						Total Score = ___ / 30

Table 1. Inter-rater agreement for the ORW

	Rater A and B		Rater A and C	
	Kappa	ICC	Kappa	ICC
Composition	.79	.89	.64	.85
Use of orthographic information	.84	.95	.67	.84
Use of letter-sound information	.74	.87	.79	.89
Use of writing vocabulary	.74	.87	1.00	1.00
Concept of word	.86	.96	.71	.89
Directionality	.62	.78	.62	.78
Rereading for information	.73	.94	.73	.94
Rereading for accuracy	.73	.92	.86	.96
Self-correcting	1.00	1.00	1.00	1.00
Fluency	.84	.95	.69	.89

$p < .05$ for all values

Table 2. Means and standard deviations of the first and final observations of ORW

		First Observation	Final Observation
	n	Mean (SD)	Mean (SD)
Section 1: Using			
Composing	24	1.75 (.53)	2.25 (.74)
Use of orthographic information	24	.63 (.58)	1.29 (.75)
Use of letter-sound information	24	1.33 (.48)	1.67 (.64)
Use of writing vocabulary	24	1.71 (.46)	1.67 (.64)
Concept of word	24	2.54 (.59)	2.75 (.44)
Direction	24	2.83 (.38)	2.96 (.20)
Sub-total Using	24	10.58 (1.74)	12.58 (2.21)
Section 2: Doing			
Rereading for information	24	1.54 (.59)	1.58 (.58)
Rereading for accuracy	24	1.33 (.64)	1.42 (.50)
Self-correcting	24	.29 (.75)	.42 (.72)
Fluency	24	1.50 (.72)	1.92 (.72)
Sub-total Doing	24	4.42 (1.72)	5.33 (1.86)
Total Score	**24**	**15.00 (3.09)**	**17.92 (3.76)**

Table 3. First observation: ORW inter-item, subtotal, total, pre- and post-writing vocabulary, pre- and post-hearing and recording sounds in words, and pre- and post-OSELA total score correlations (n = 24)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. COMP	-																		
2. OI	.25	-																	
3. LSI	-.92	.38	-																
4. WV	-.70	.42	.25	-															
5. COW	-.14	.19	.33	.13	-														
6. DIR	.06	.42	.22	.17	.33	-													
7. RRS	.21	.49	.25	.40	.51	.28	-												
8. RRA	.29	.54	.17	.31	.13	.46	.62	-											
9. SC	-.17	.08	.09	.41	-.21	.15	.21	.11	-										
10. FL	-.13	.16	.42	.27	.31	.41	.52	.32	.38	-									
11. USE	.31	.79	.59	.59	.55	.57	.65	.55	.09	.42	-								
12. DO	.71	.43	.37	.31	.43	.47	.82	.72	.54	.81	.60	-							
13. EWR	.21	.68	.54	.51	.55	.58	.82	.70	.35	.68	.90	.89	-						
14. EWV	.28	.10	.32	-.05	-.05	.25	-.03	.07	.05	-.02	.24	.12	.21	-					
15. XWV	-.24	.19	.39	.07	.20	-.09	.14	.17	.32	.45	.21	.41	.30	.37	-				
16. EHR	.24	.12	.26	.08	.10	.18	.07	.07	.21	.39	.27	.27	.32	.69	.37	-			
17. XHR	-.28	.33	.28	.08	.07	.29	-.01	.04	.25	.16	.17	.21	.20	.14	.55	.23	-		
18. EOS	.27	.25	.15	.07	-.04	.20	-.12	.07	.13	.43	.27	.28	.30	.89	.44	.89	.25	-	
19. XOS	-.18	.45	.33	.12	.09	.05	.38	.19	.25	.31	.30	.33	.32	.42	.89	.42	.77	.51	-

Note: Correlations larger than .39 are statistically significant, $p < .05$.
COMP = Composition; OI = Use of Orthographic Information; LSI = Use of letter sound information; WV= Writing vocabulary; COW = Concept of word; DIRECT = Concepts about directionality; RRS = Rereading as if to search; RRA = Rereading for accuracy; SC = Self-correcting; FL = Fluency; USE = Using sub-total score; DO = Doing sub-total score; EWR = EWOR Total Score; EWV = Pre-intervention Writing Vocabulary; XWV = Post-intervention Writing Vocabulary; EHR = Pre-intervention Hearing and Recording Sounds in Words; XHR = Post-intervention Hearing and Recording Sounds in Words; EOS= Pre-intervention OSELA total score; XO = Post-intervention OSELA total score.

Table 4. Final observation: ORW inter-item, subtotal, total, pre- and post-writing vocabulary, pre- and post-hearing and recording sounds in words, and pre- and post-OSELA total score correlations (*n* = 24)

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. COMP	-																		
2. OI	.41	-																	
3. LSI	.19	.49	-																
4. WV	.46	.30	.14	-															
5. COW	.33	.49	.15	.00	-														
6. DIR	.07	.08	.22	-.11	.36	-													
7. RRS	.46	.49	.43	.43	.08	.21	-												
8. RRA	.29	.35	.45	.18	.29	.18	.62	-											
9. SC	.12	.33	.22	.22	.20	.12	.23	.22	-										
10. FL	.62	.53	.32	.32	.48	.27	.54	.46	.32	-									
11. USE	.74	.81	.61	.58	.56	.25	.62	.47	.33	.69	-								
12. DO	.51	.58	.47	.39	.37	.27	.78	.73	.64	.80	.72	-							
13. EWR	.68	.76	.58	.53	.51	.28	.74	.64	.51	.80	.91	.94	-						
14. EWV	-.11	.20	.16	-.21	.28	.12	.05	-.21	-.21	.11	.24	.12	.37	-					
15. XWV	.13	.47	.38	-.03	.34	.25	.35	.32	.58	.42	.32	.48	.47	.37	-				
16. EHR	-.11	.19	.15	-.08	.15	-.22	-.11	-.04	.08	.07	.07	.06	.05	.69	.37	-			
17. XHR	.04	.35	.25	.06	.26	-.09	.31	.51	.26	.16	.19	.35	.30	.20	.55	.23	-		
18. EOS	-.07	.29	.12	-.15	.20	-.06	-.01	-.17	.06	.11	.12	.05	.10	.87	.44	.89	.25	-	
19. XOS	.08	.51	.44	-.05	.28	.10	.38	.39	.34	.33	.32	.42	.43	.42	.89	.42	.77	.50	-

Table 5. Linear growth models of ORW scores (unconditional model)

Using Sub-total				
Fixed Effect	*Coefficient*	*se*	*t Ratio*	*p value*
Mean initial status, β_{00}	10.58	.35	30.35	<.001
Mean growth rate, β_{10}	1.74	.33	5.28	<.001
Random Effect	*Variance Component*	*df*	X^2	*p value*
Initial Status, r_{00}	1.86	23	40.14	<.01
Growth rate, r_{1i}	.31	23	15.68	>.5
Level-1 error, e_{1i}	1.84			
Reliability of OLS Regression Coefficient Estimate				
Initial Status, π_{01}	.51			
Growth Rate, π_{1i}	.04			
Doing Sub-total				
Fixed Effect	*Coefficient*	*se*	*t Ratio*	*p value*
Mean initial status, β_{00}	4.46	.38	11.86	<.001
Mean growth rate, β_{10}	.76	.31	2.47	<.05
Random Effect	*Variance Component*	*df*	X^2	*p value*
Initial Status, r_{00}	1.86	23	40.14	<.01
Growth rate, r_{1i}	.31	23	15.68	>.5
Level-1 error, e_{1i}	1.84			
Reliability of OLS Regression Coefficient Estimate				
Initial Status, π_{01}	.51			
Growth Rate, π_{1i}	.04			
EWOR Total Score				
Fixed Effect	*Coefficient*	*se*	*t Ratio*	*p value*
Mean initial status, β_{00}	15.03	.65	23.24	<.001
Mean growth rate, β_{10}	2.54	.56	4.53	<.001
Random Effect	*Variance Component*	*df*	X^2	*p value*
Initial Status, r_{00}	5.39	23	40.85	<.01
Growth rate, r_{1i}	.54	23	20.42	>.5
Level-1 error, e_{1i}	5.50			
Reliability of OLS Regression Coefficient Estimate				
Initial Status, π_{01}	.50			
Growth Rate, π_{1i}	.07			